Metody probabilistyczne i statystyka

Estymacja i testowanie hipotez statystycznych dla jednej populacji w R[a]

Szybkie wprowadzenie do statystyki[b]

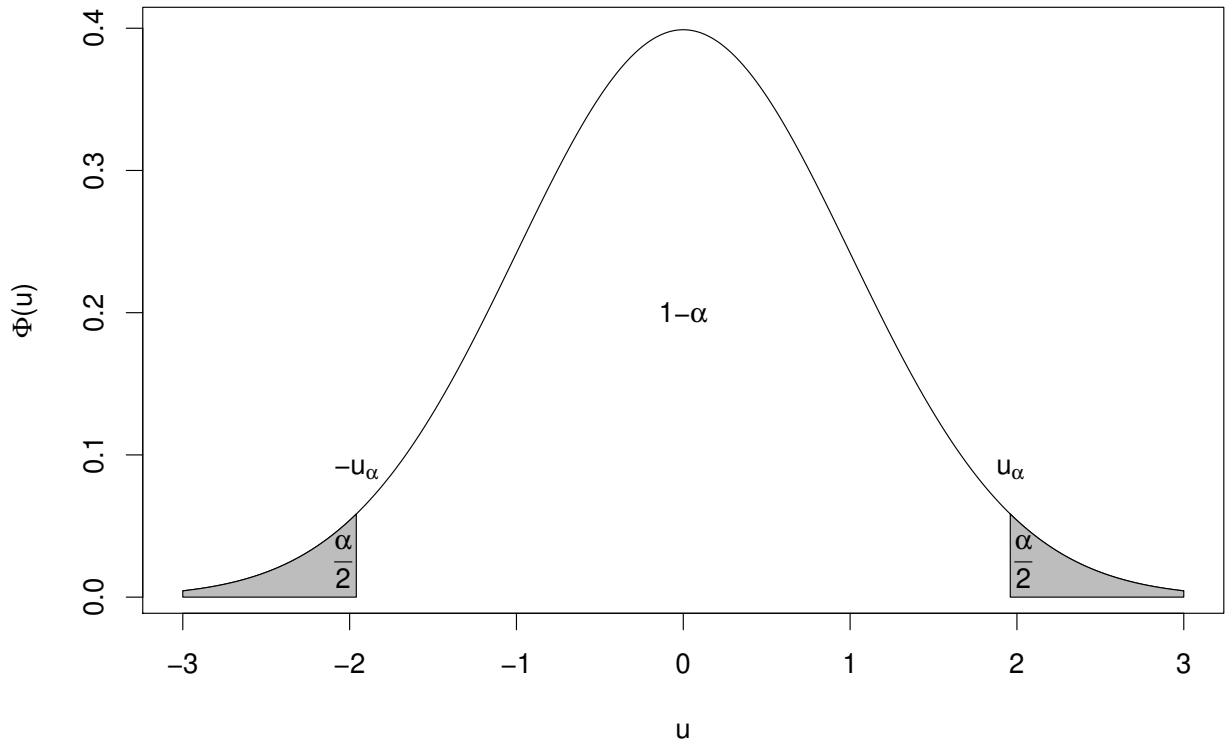
— Jeżeli populacja generalna ma rozkład normalny i znane jest odchylenie standardowe σ, to przedział ufności dla średniej otrzymuje się ze wzoru:

$$P\left(\overline{x} - u_\alpha \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + u_\alpha \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha, \tag{1}$$

gdzie $\overline{x}$ oznacza średnią arytmetyczną z próby, $1-\alpha$ jest prawdopodobieństwem przyjętym z góry i nazywanym *współczynnikiem ufności*, a u_α jest wartością zmiennej losowej U mającej rozkład normalny standaryzowany. Wartość u_α dla danego współczynnika ufności $1-\alpha$ wyznacza się w taki sposób, by spełniona była relacja

$$P\left(-u_\alpha < U < u_\alpha\right) \equiv P\left(|U| < u_\alpha\right) = 1 - \alpha. \tag{2}$$

Na przykład dla najczęściej przyjmowanego przedziału ufności $1 - \alpha = 0{,}95$ mamy $u_\alpha \approx 1{,}96$.



Rysunek 1. Rozkład normalny standaryzowany.

[a] Dla wszystkich testów podać hipotezy, wartości statystyk, poziom p i wnioski.

[b] Tekst pochodzi w dużej mierze z książki STATYSTYKA MATEMATYCZNA MODELE I ZADANIA autorstwa Jerzego Grenia.

— Jeżeli populacja generalna ma rozkład normalny z nieznanym odchyleniem standardowym, to przedział ufności dla średniej otrzymuje się ze wzoru:

$$P\left(\overline{x} - t_\alpha \cdot \frac{s}{\sqrt{n-1}} < \mu < \overline{x} + t_\alpha \cdot \frac{s}{\sqrt{n-1}}\right) = 1 - \alpha, \tag{3}$$

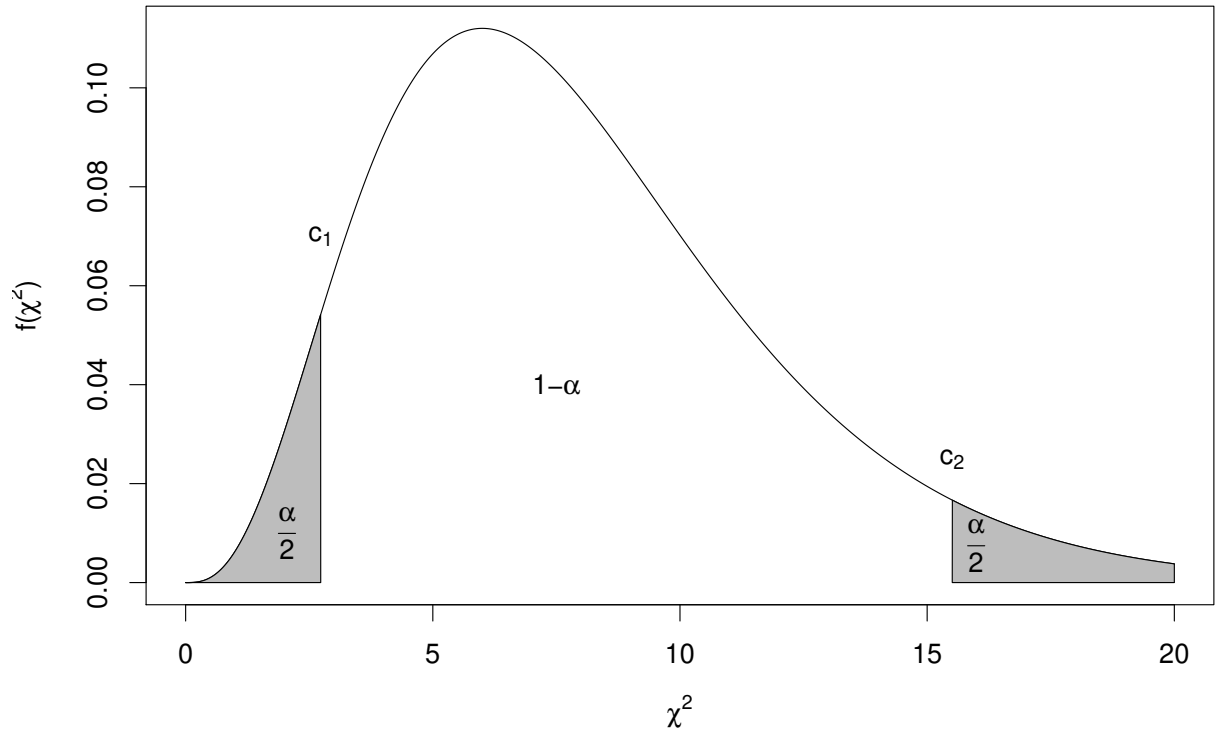
gdzie $\overline{x}$ oznacza średnią arytmetyczną z próby, a s oznacza odchylenie standardowe z próby. Wartość t_α oznacza wartość zmiennej t Studenta z $n-1$ stopniami swobody, która dla danego z góry prawdopodobieństwa $1-\alpha$ spełnia relację

$$P\left(-t_\alpha < t < t_\alpha\right) \equiv P\left(|t| < t_\alpha\right) = 1 - \alpha. \tag{4}$$

— Jeżeli populacja generalna ma rozkład normalny o nieznanych parametrach μ i σ i wylosowana próba jest nieduża ($n < 30$), to przedział ufności dla wariancji σ^2 populacji generalnej określa wzór

$$P\left(\frac{ns^2}{c_2} < \sigma^2 < \frac{ns^2}{c_1}\right) = 1 - \alpha, \tag{5}$$

gdzie c_1 i c_2 są wartościami zmiennej χ^2 z $n-1$ stopniami swobody wyznaczonymi w taki sposób, by spełnione były relacje $P\left(\chi^2 < c_1\right) = \frac{\alpha}{2}$ oraz $P\left(\chi^2 \geqslant c_2\right) = \frac{\alpha}{2}$.



Rysunek 2. Rozkład χ^2.

W przypadku dużej próby przybliżony przedział ufności dla odchylenia standardowego σ populacji generalnej jest określony wzorem (u_α wyznacza się tak samo jak w (2))

$$P\left(\frac{s}{1 + \frac{u_\alpha}{\sqrt{2n}}} < \sigma < \frac{s}{1 - \frac{u_\alpha}{\sqrt{2n}}}\right) \approx 1 - \alpha. \tag{6}$$

— Do weryfikacji *hipotezy statystycznej* służy *test statystyczny* — ograniczymy się jedynie do *testu istotności*. Na jego potrzeby formułuję się *hipotezę zerową* H_0 (bezpośrednio sprawdzaną, np. „*wartość przeciętna wynosi 100 cm*", „*wariancje w obu populacjach są równe*") oraz hipotezę alternatywną H_1. Jednym z parametrów testu istotności jest *poziom istotności* α, która oznacza prawdopodobieństwo popełnienia *błędu pierwszego rodzaju* (błąd polega na odrzuceniu prawdziwej hipotezy H_0).

— Jeśli poziom $p < \alpha$ (wartość statystyki znajduje się w obszarze krytycznym), to hipotezę zerową się odrzuca; w przeciwnym wypadku nie ma podstaw do odrzucenia H_0.

Szybkie wprowadzenie do `R`

— Do zweryfikowania, czy próba pochodzi z rozkładu normalnego, posłużą się Państwo *testem Shapiro-Wilka* — w `R` służy do tego funkcja `shapiro.test`, do której należy przekazać wektor liczb. Funkcja zwraca obiekt klasy `htest`, która zawiera pola: `statistic` (wartość statystyki oznaczoną literą `W` w podsumowaniu), `p.value` (poziom p), `method` (zawiera napis „`Shapiro-Wilk normality test`") oraz `data.name` (nazwa parametru).

— Do przeprowadzenia testu dla wartości przeciętnej w jednej populacji posłużą się Państwo funkcją `t.test`, która przyjmuje następujące parametry:

⋆ `x` — wektor wartości (jest to pierwszy parametr funkcji),

⋆ `alternative` — określa obszar krytyczny: dopuszczalne wartości to `"two.sided"` (test z obustronnym obszarem krytycznym — hipoteza alternatywna ma postać $m \neq$ `mu`), `"less"` (test z lewostronnym obszarem krytycznym — hipoteza alternatywna ma postać $m <$ `mu`), `"greater"` (test z prawostronnym obszarem krytycznym — hipoteza alternatywna ma postać $m >$ `mu`),

⋆ `mu` — testowana wartość średniej,

⋆ `conf.level` — poziom ufności.

Funkcja zwraca obiekt klasy `htest`, zawierający m.in. te same pola, co w przypadku funkcji `shapiro.test`.

— Dla przypomnienia: blok decyzyjny ma postać:

```
if (warunek) {
    # gdy 'warunek == TRUE'
} else {
    # gdy 'warunek == FALSE'
}
```

Zadanie

Dla zmiennej `Wzrost` w grupie mężczyzn:

— dokonać wstępnej oceny zgodności z rozkładem normalnym na podstawie wykresu *„kwantyl-kwantyl"* (funkcja `qqnorm` sporządzi wykres, a funkcja `qqline` wyznaczy prostą przechodzącą przez dolny i górny kwartyl — proszę ocenić skupienie punktów wokół prostej),

— na poziomie istotności $\alpha = 0{,}01$ sprawdzić założenie o normalności rozkładu testem Shapiro-Wilka,
— na poziomie istotności $\alpha = 0{,}05$ zweryfikować hipotezę, że średni wzrost w populacji jest większy niż 179 cm,
— wyznaczyć przedziały ufności dla średniej (napisz w tym celu funkcję) wzrostu w populacji (poziomy ufności 0,95 oraz 0,98)[c] — jak poziom ufności wpływa na szerokość przedziału ufności?
— wyznaczyć przedział ufności dla odchylenia standardowego (napisz w tym celu funkcję) wzrostu w populacji (poziom ufności 0,97)[d]

Pod koniec zajęć proszę przysłać skrypt na adres `pt.labki@gmail.com`, wpisując w temacie numer grupy.

[c] Odp. [178,8865; 181,7728] dla poziomu ufności 0,95.
[d] Odp. [5,962915; 8,256527] (używając (5)) albo [5,936392; 8,212323] (używając (6)).